Defense Technical Information Center Compilation Part Notice

ADP014027

TITLE: Large Vocabulary Audio-Visual Speech Recognition

DISTRIBUTION: Approved for public release, distribution unlimited

This paper is part of the following report:

TITLE: Multi-modal Speech Recognition Workshop 2002

To order the complete compilation report, use: ADA415344

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, etc. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:
ADP014015 thru ADP014027

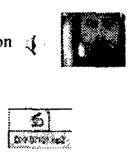
Large Vocabulary Audio-Visual Speech Recognition

Chalapathy Neti & Gerasimos Potamianos
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

- •Motivation
- •A/V speech recognition architecture
- •Visual feature extraction
- •Audio-visual fusion
- •Results
- •Challenges and conclusions

Pervasive Speech recognition

- ➢ Pervasive deployment of speech will require better recognition in degraded acoustic conditions:
  - High noise ("speech babble") e.g.
    - ✓ Military applications
    - ✓ Automobiles
    - ✓ Video Games & Interactive television
  - Whispered Speech
    - Privacy
  - Lombard speech
    - High-noise conditions
  - Speech pathology

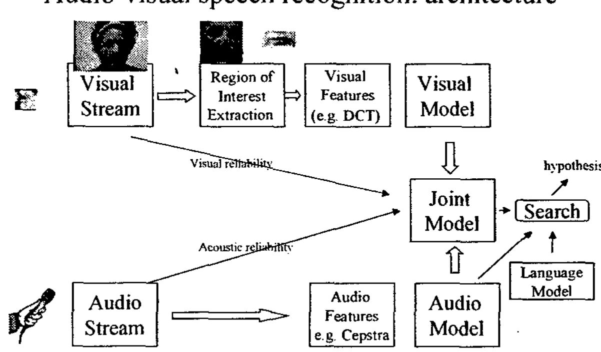


Audio-Visual speech recognition is a key enabler

IBM's A/V speech effort

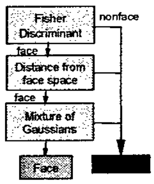
- History (www.research.ibm.com/AVSTG)
  - About a 3 year old effort.
  - Led the JHU Workshop team on A/V speech recognition, 2000.
  - AVSTG department formed in 2001
  - Taught an invited ELSNET tutorial on A/V speech recognition (Prague, 2001).
- Highlights/differentiators of our work
  - One of a kind database for AV LVCSR
  - State-of-the-art audio ASR subsystem (LVCSR)
  - Fully automated visual front end
    - Multiresolution face detection
    - Augmented visual speech ROI (jaw region instead of mouth)
    - Multistage (linear transform based) visual feature extraction
  - Sub-phonetic visual speech models
    - Scales to large-vocabulary recognition
  - Phone-level asynchronous A/V fusion
    - Joint a/v model training
    - Maximum entropy based stream weight estimation (global and local)
  - Multiple domain exploration
    - Read speech (digits/C&C/LVCSR), Impaired speech, Automobile, Broadcast News
    - Visual adaptation to new domains

Audio-visual speech recognition: architecture



Visual Front-End

- Multiresolution face detection:
  - Search for skin-tone pixels
  - Search image pyramid across scales & locations.
  - Each square *mxm* region is considered as a face.
  - Hierarchical, pixel based approach, using LDA and PCA.



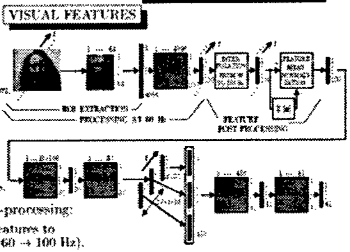
5-level pyramid with face detector at each level

ROI EXTRACTION

- Steps (after face/mouth tracking):
  - *Smooth* mouth **center** and **size** estimates by *median* filtering.
  - Extract a 64 × 64 pixel, *size-normalized* mouth ROI.
  - ROI includes jaw and cheeks.

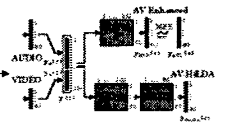
VISUAL FEATURES

- Three stages:
  - I: *Discrete cosine transform* **(DCT)**, for data *compression*.
  - II: *Intra-frame* **LDA/MLLT**: 100 → 30 *static* features.
  - III: *Inter-frame* **LDA/MLLT**: 15 × 30 → 41 *dynamic* feats.
- DCT feature post-processing:
  - *Interpolate* visual features to audio feature rate (60 → 100 Hz).
  - Apply *feature mean normalization* for lighting *compensation*.

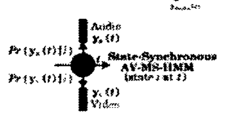


Audio-visual Fusion Techniques

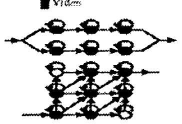
- Feature fusion
  - Enhancement approach
  - Discriminant fusion (HiLDA)



- State Synchronous Multistream HMM
  - Allows weighting decisions



- Phone synchronous Multistream HMM
  - Allows asynchrony within a phone



THE MULTI-STREAM HMM FOR AV-ASR

- The multi-stream (MS) HMM:
  - Observation conditional "score" of audio-visual state $i \to (i_A, i_V)$:

$$\mathcal{L}(\mathbf{y}(t)|i) = \left[Pr(\mathbf{y}_A(t)|i_A)\right]^{\lambda_A} \times \left[Pr(\mathbf{y}_V(t)|i_V)\right]^{\lambda_V}$$

  - *Exponents* model stream "reliability". Typically:

$$0 \leq \lambda_A, \lambda_V \leq 1 \,,\; \lambda_A + \lambda_V = 1$$

- State vs. phone level synchronous (product) MS-HMM:
  - **State synchrony**: $\{i_A\} = \{i_V\}$, $i = i_A = i_V$.
  - **State asynchrony**: $i \in \{i_A\} \times \{i_V\}$.
- MS-HMM parameter training:
  - **Model parameters**: $\underline{\theta} = [\underline{\theta}_A, \underline{\theta}_V, \lambda_A, \lambda_V]$, where $\underline{\theta}_A$, $\underline{\theta}_V$ are audio- or visual-only HMM stream params.
  - *Maximum likelihood* estimation (MLE) of $\underline{\theta}_A$, $\underline{\theta}_V$ via **EM**:
    - *Independent* **E-** and **M-steps** for MLEs of $\underline{\theta}_A$, $\underline{\theta}_V$.
    - *Joint* audio-visual MS-HMM **E-step**. **M-step**, as above.
  - MLE of λ_A, λ_V is impossible. Instead, we have considered:
    - *Parameter grid search*. Minimizes held-out data **WER**.
    - *Minimum classification error* **(MCE)** training by **GPD**.
    - *Maximum entropy*. Maximizes data posterior log-likelihood.

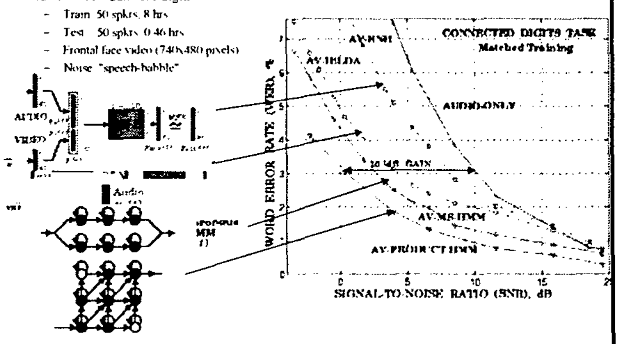
IBM VVAV databases

- LVCSR
  - *First-of-a-kind audio-visual database for large-vocabulary continuous SI speech recognition (LVCSR)*
  - 290 subjects
  - 70 hrs. continuous speech, 10,400 word vocabulary
- Digits
  - 50 subjects
  - 8.46 hrs. continuous speech, 11 word vocabulary
- Database Format
  - Frontal face color video, 704x480, 30 Hz, MPEG2
  - 16 kHz/16bit pcm

Experiments on Digits

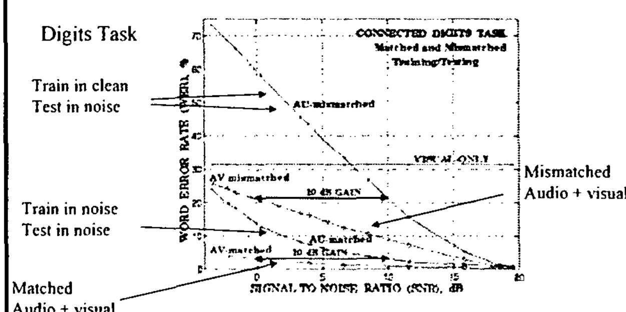
Fusion Techniques

- IBM VVAV database digits
  - Train 50 spkrs, 8 hrs
  - Test 50 spkrs 0.46 hrs
  - Frontal face video (740x480 pixels)
  - Noise "speech-babble"



Results - Summary

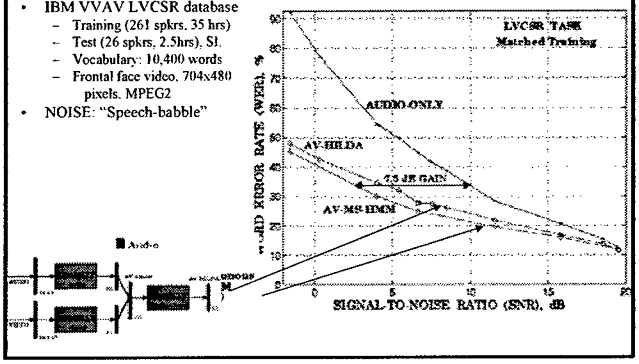
- Effective gain of 10 dB @ 10 dB SNR (relative to mismatched audio)
- Effective gain of 10 dB @ 10 dB SNR (relative to matched audio)



Experiments on LVCSR

Results: LVCSR

- IBM VVAV LVCSR database
  - Training (261 spkrs. 35 hrs)
  - Test (26 spkrs. 2.5hrs). SI.
  - Vocabulary: 10,400 words
  - Frontal face video. 704x480 pixels. MPEG2
- NOISE: "Speech-babble"



Results: Human .vs. Machine

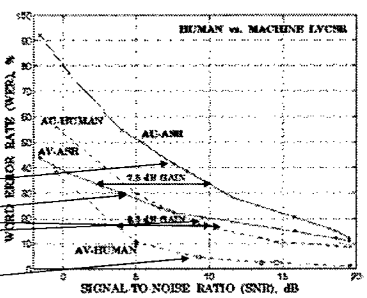
- IBM VVAV LVCSR database
  - Training (261 spkrs. 35 hrs)
  - Test (26 spkrs. 2.5hrs). SI.
  - Vocabulary: 10,400 words
  - Frontal face video. 704x480 pixels. MPEG2
- NOISE: "Speech-babble"

Audio-ASR

AV-ASR

Audio- human

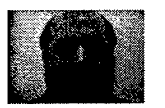
AV-human



Super-human performance below 7dB SNR

Challenges

- IBM VVAV data
  - Audio
    - Read Speech, single microphone
    - Additive "speech babble" noise
  - Video
    - Frontal Face, uniform background
    - Uniform lighting

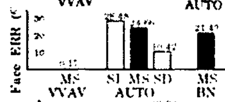


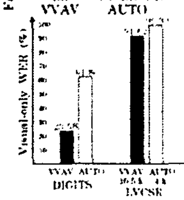
- Broadcast Video data
  - Audio
    - Spontaneous speech
    - Additive music, varying channel, etc.
  - Video
    - Limited pose variation, background clutter
    - Uniform lighting

- Automobile data
  - Audio
    - Spontaneous speech
    - Automobile noise (speed variation, radio, seat belt, etc.)
  - Video
    - Pose variation, varying background
    - Non-uniform lighting Conditions

VISUALLY CHALLENGING DOMAINS: PRELIMINARY RESULTS

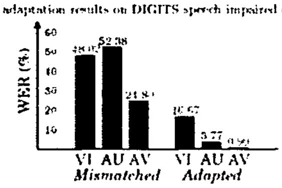
- Challenge: Video data variability in *head pose, background, and lighting* affects *face detection*, ROI *extraction/normalization*, thus visual- and AV-ASR.
- Face detection error for VVAV, AUTO (multi-speaker vs speaker-(in)dependent) and BN data
- Visual-only WER for VVAV vs AUTO domains (DIGITS and LVCSR tasks)





AUDIO-VISUAL SPEAKER ADAPTATION

- Important for speaker enrollment and limited data domains, but *hardly* ever considered in the AV-ASR literature
- Main techniques:
  - MLLR: Rapid adaptation of HMM stream component *means*
  - MAP: Bayesian approach, adapts *all* HMM parameters.
  - FE. Front end adaptation of *LDA/MLLT matrices*.
- The domains/tasks considered:
  - Domains: Noisy *trading* floor; hearing *impaired* speech
  - Tasks: LVCSR; DIGITS.
- MLLR adaptation results on DIGITS speech impaired data:



Conclusions

- Consistent and significant gains for all audio conditions
- Significant performance gains in "speech-babble" noise
  - Effective gain of 10 dB @ 10 dB SNR for digits
  - Effective gains of 7.5 dB @ 10 dB for LVCSR
- Significant gains in relatively clean environments
  - 62% relative gain for digits (0.75 -> 0.28)
  - 8% for LVCSR
- Super-human performance at high-noise levels
- Asynchrony modeling helps for digits
- Further research required in visually challenging domains
- Visual adaptation is a promising approach
  - Upto 67% relative improvement in visual speech recognition

Who ? What ? Where ? How ?

Perceptually Aware User Interfaces

Alex Waibel

June 11, 2002
Interactive Systems Laboratories
Carnegie Mellon University
University of Karlsruhe

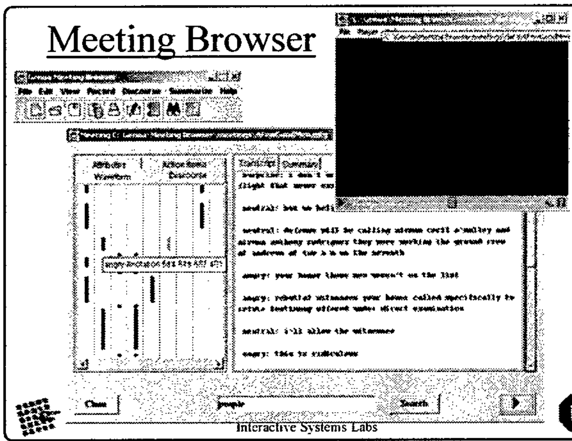
http://www.is.cs.cmu.edu
Email: waibel@cs.cmu.edu

Interactive Systems Labs

Meetings

Interactive Systems Labs

Meeting Browser



Interactive Systems Labs

Interpreting Human Communication

"Why did Joe get angry at Bob about the budget ?"

Need Recognition and Understanding of Multimodal Cues

- Verbal:
  - Speech
    - Words
    - Speakers
    - Emotion
    - Genre
  - Language
  - Summaries
  - Topic
  - Handwriting
- Visual
  - Identity
  - Gestures
  - Body-language
  - Track Face, Gaze, Pose
  - Facial Expressions
  - Focus of Attention

Interactive Systems Labs

Human Interaction

- People ID – Who?
  - Speaker ID, Face ID
  - Type: Dominant, Submissive, etc.
  - Relationship: Family, Friends, Colleague
- Speech and Discourse – What ?
  - Speech: Transcript
  - Discourse States (Speech Acts, Topics), Games, Turn Taking
  - Discourse Types and Genres (Negotiation, Chatting, Lecturing)
- Emotional State, Affect – How ?
  - Angry, Happy, Sad, Afraid;... Busy, Nervous, Relaxed
  - Discourse Style: Sloppy, Formal, Colloquial
- Localization and Speaker and Focus of Attention – Where ?
  - Speaker Localization
  - Focus of Attention Tracking

Interactive Systems Labs

Main Challenge and Goal

Robustness in Real-Life Situations

Interactive Systems Labs

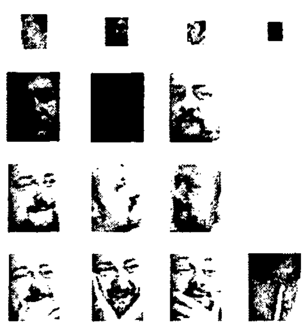
Visual Challenges

Low quality

Illumination

Head pose

Occlusion



Interactive Systems Labs

Acoustic Challenges

- Sloppy Speech
- Noise
- Reverberation
- Acoustic Scene Analysis
- Cross Talk
- Distant Mic

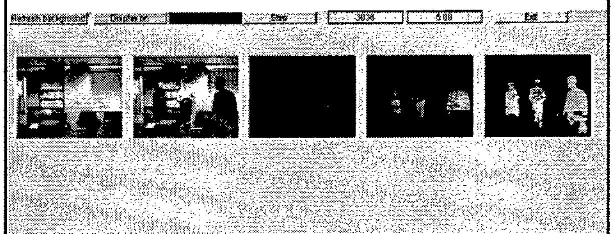
Interactive Systems Labs

Where ?

- Face Tracking (Visual)
- Sound Source Localization (Acoustic)
- People Tracking (Visual)
- Behavior and Movement Models

Interactive Systems Labs

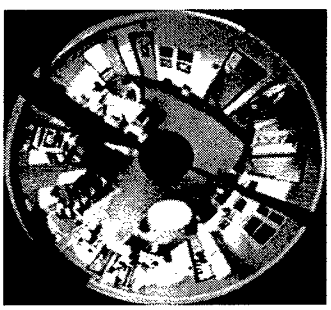
Tracking People



Tracking People

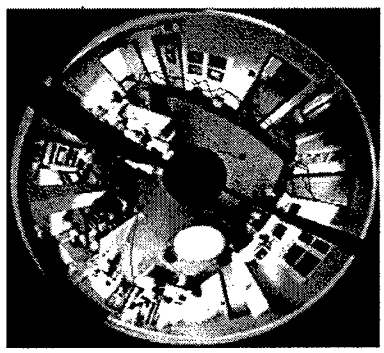
Interactive Systems Labs

Tracking Multiple People



Interactive Systems Labs

From Tracking to Modeling Activity



Interactive Systems Labs

Real-Time Face Tracker

Three Types of Models have been employed

- skin-color model to register the face
- motion model to estimate image motion
- camera model to predict and compensate for camera motion (pan, tilt, zoom)

The Face Tracker

- tracks a persons face while person is freely moving (e.g. walks, jumps, sits down and stand up)
- Framerate : 30+ frames per second using a low end workstation (HP9000) or Pentium II 266 PC.

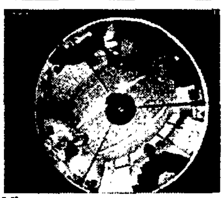
Interactive Systems Labs

Real-Time Face Tracker

Interactive Systems Labs

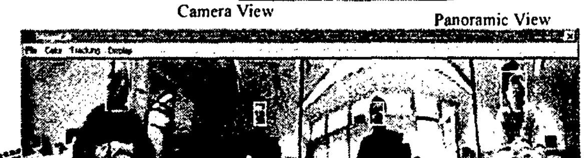
Using a Panoramic Camera

Cyclovision's ParaCam



Camera View

Panoramic View



Interactive Systems Labs

Pose Tracking by Modeling Shape

Interactive Systems Labs

Face and Pose Tracking

FaceTracking

Pan-Tilt-Zoom

Panorama

GazeTracking

Interactive Systems Labs

What ?

- Large Vocabulary Speech Recognition
  - Issues:
    - Sloppy Speech
    - Distant Microphones
    - Mismatch in Vocabulary
    - Other Languages
  - Many Other Aspects: Topic Detection, Named Entity, Translation, Discourse, ...
- Multimodal Dialog
  - Fuse Speech, Pointing, Gesture, Handwriting
  - Fusion Usually at Semantic Level
- Audio-Visual Speech
  - Combine Speech and Visual Info

Interactive Systems Labs

From Read Speech to Conversational Speech

Conversational Speech

- Wall Street Journal Dictation
- Broadcast News Database
  - Transcription and Information Retrieval on News Casts
  - Multilingual Speech Recognition
- Switchboard & Callhome
  - Human to Human Telephone Speech
- Meetings and Discussion Database
  - Newshour (18h)
  - Crossfire (9h)
  - Group Meetings

Interactive Systems Labs

Transcribing Speech in Meetings

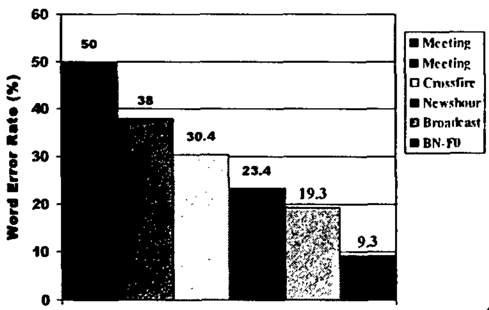
Run-On Transcription of Meetings

- *Mismatched* Recording Conditions
  - Remote Microphones
  - Cross-Talk
  - Recording Always on !
  - Noise
  - Multiple Speakers
- *Mismatched* Speaking Style:
  - **Spontaneous and Conversational** Human to Human Speech
  - Emotional Speech
- *Mismatched* Language and Vocabulary
  - Special Ideosynchratic Topic

- Three Tasks:
  - Newshour
  - Crossfire
  - Group Meetings

Interactive Systems Labs

Recognition of Conversational Speech

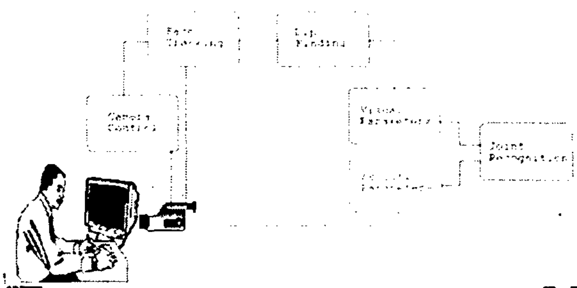


Interactive Systems Labs

Audio-Visual Speech:

(When Acoustic Processing is not Good Enough)

- Duchnowski, Manke, Bregler, Meier, Waibel
  - ICASSP'93, ICSLP'94, ICASSP'95, ...



Interactive Systems Labs

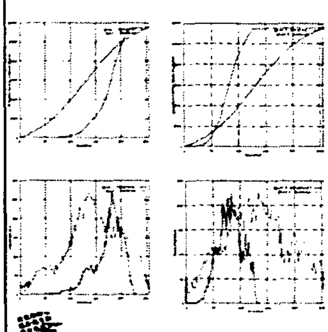
Visual Preprocessing

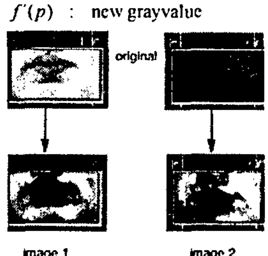
grayvalue modification - example histogram : $f'(p) = T(f(p))$

$f(p)$: original grayvalue

T : modification function

$f'(p)$: new grayvalue



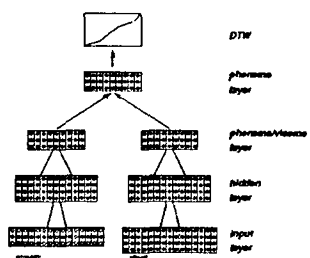


Interactive Systems Labs

Audio-Visual Recognizer

$hyp_c = \lambda_a\, hyp_a + \lambda_v\, hyp_v$

$1 = \lambda_a + \lambda_v$



Features
- What Features to Use?

Fusion Level
- Feature Vector
- Phone Streams
- Word Level

Fusion Methods
- Trained Weights
- Entropy Weights
- SNR Weights

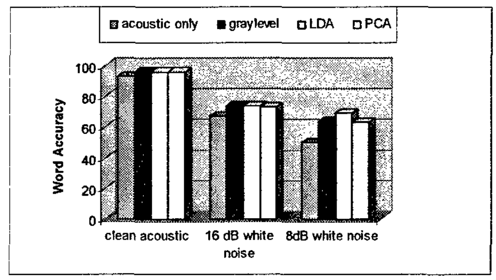
Interactive Systems Labs

Experiments

- Task: Continuous Letter Spelling
  - Difficult, but smaller Vocabulary
- Speaker dependent
  audio-visual results
  - Fusion by Entropy Weights
  - LDA Front End
  - Phone Level Fusion

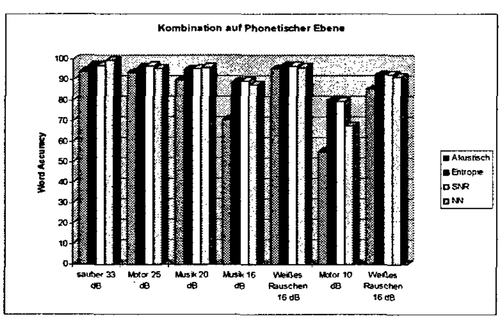
Interactive Systems Labs

Audio-Visual Fusion



Interactive Systems Labs

Fusion Weights and Noise



Interactive Systems Labs

Who ?

- Once we have found the Face
- Face ID
  - Problems: Occlusion
- Speaker ID
  - Problems: Distant Mics, Noise, Jamming Noise
  - Phonetic Speaker ID, Qin Jing

Interactive Systems Labs

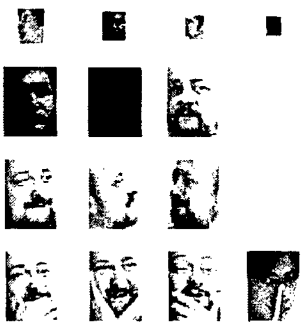
People Identification: Challenges

Low quality

Illumination

Head pose

Occlusion



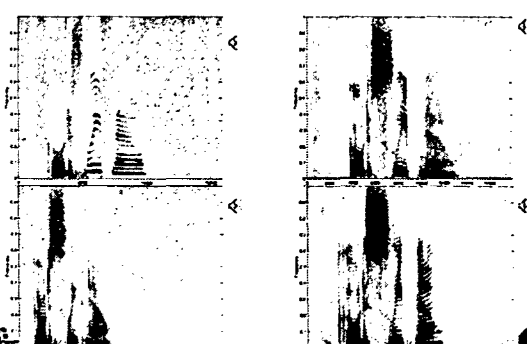
Interactive Systems Labs

How ?

- Detect Emotional State
  - Happy, Angry, Sad, Afraid
  - Distress, Busy, Relaxed...
- Techniques:
  - Acoustic: (Polzin, 1999)
    - Prosody: Intensity, Pitch, Rhythm,
    - Language: Words and Expressions Used
  - Visual: (Cohen)
    - Facial Expressions

Interactive Systems Labs

Emotion: Acoustic Information



Interactive Systems Labs

Emotion: Language Information

- Lexical metaphors

  Son of a bitch!
  (As Good as it Gets)
- Connotation-loaded lexems

  You're a spoiled rotten little brat!
  (Kramer versus Kramer)
- Intensification

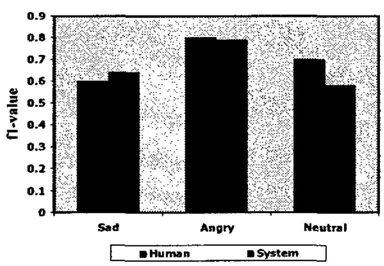
  We ain't got the slightest f... idea what happened ...
  (Reservoir Dogs)

  That makes me very very mad ...
  (The Sweet Hereafter)

Interactive Systems Labs

Performance Comparison (Movies)

Verbal and Non-Verbal Information



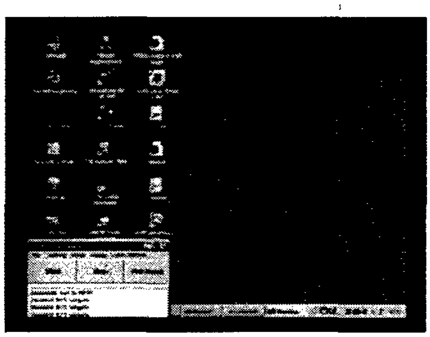
Interactive Systems Labs

To Whom ?

- Focus of Attention Tracking
  - R. Stiefelhagen, PUI'98, Humanoids'01, PhD Thesis'02
  - Who is addressee of an utterance ?
  - Who is someone making talking to ?
  - What is a human user attending to ?
- Observation:
  - FoA is a Psychological State, can only be infered or 'guessed' from correlates
  - Both Observed User and Target are important:
    - Pose, Eye-Gaze
    - Possible Targets: Noise, Movement, Faces, Speech

Interactive Systems Labs

Focus of Attention Tracking



Interactive Systems Labs

Conclusion

- Complete Model of Human Communication is Needed
  - Include all modalities
  - Include different not only *what* was said, but also: *who, where, to whom, how…*
- Challenges:
  - *Robust* Processing of Component
  - Proper Level and Method of Fusion
  - Robust and *Dynamic* Fusion of Useful Clues

Interactive Systems Labs

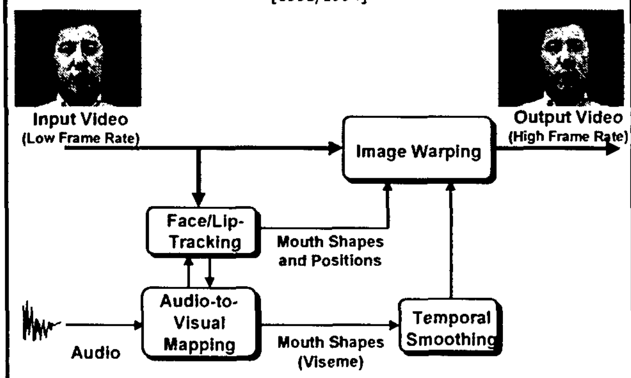
Joint Audio-Visual Speech Recognition and CMU Audio-Visual Speech Data Set

Prof. Tsuhan Chen
Carnegie Mellon University

Thanks to Dr. Simon Lucey and Jie Huang

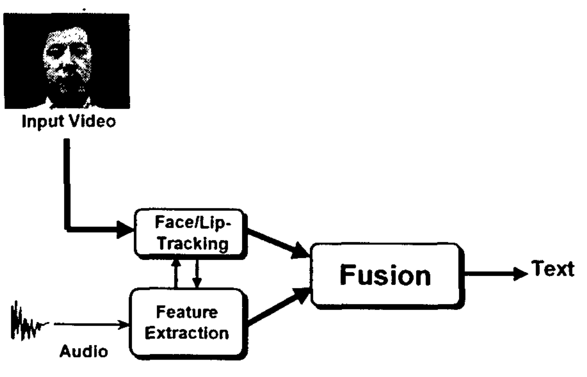
Where We Started...

[1993/1994]



Tsuhan Chen

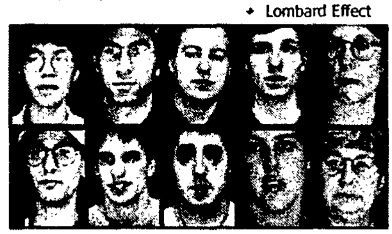
Lip-Reading



Tsuhan Chen

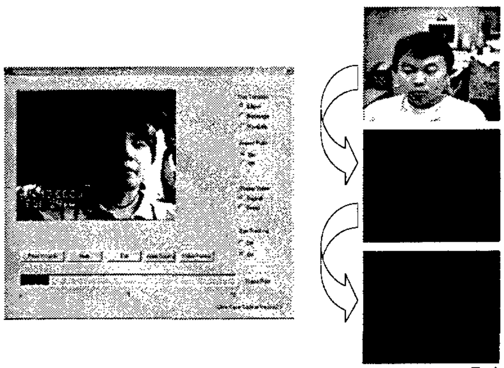
Audio-Visual Speech Data Set

- Thanks to Intel
- 78 isolated words 10 times
  - Date/time/month/day/etc.
  - Audio: 44.8 kHz, 16 bits
  - Video: 30/60Hz, 720×640
- Lip parameters extracted
- Noises
  - Gaussian white/pink noise, car, factory (Noise-X 92)
  - Babble/crosstalk
  - Lombard Effect



Tsuhan Chen

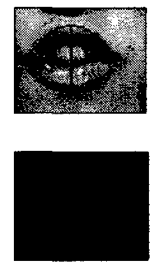
Face Tracking

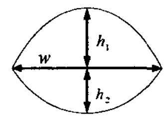


Tsuhan Chen

Lip Tracking

- Modeling color distribution of mouth pixels
  - Gaussian mixture
- Deformable template





Tsuhan Chen

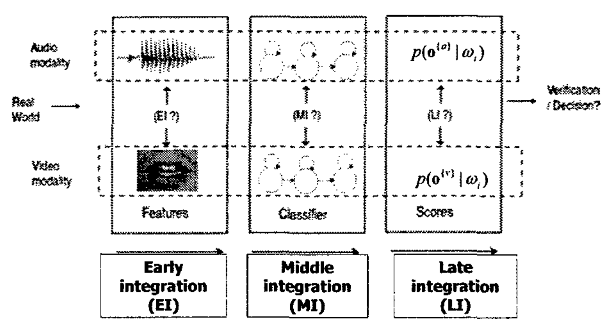
Customers...

- "Signal Processing for Media Integration," ICASSP 2002
  - Coupled HMM for Audio-Visual Speech Recognition, Nefian et al., Intel
  - Visual Speech Feature Extraction for Improved Speech Recognition, Zhang, Mersereau, Clements, Georgia Tech
  - Audio-Visual Speech Modeling Using Coupled HMM, Chu, Huang, UIUC
- Others

  California State University
  Chunghwa Telecom Lab, Taiwan
  DongYang University, Korea
  Fabbrica Servizi Telematici, Italy
  IIT Bombay, India
  Instituto Tecnologico de Buenos Aires
  On2.com
  Queensland University of Technology
  National Tsinghua University
  National University of Singapore
  Norwegian Computing Center, Norway
  Shanghai JiaoTong University, China
  Washington University

Tsuhan Chen

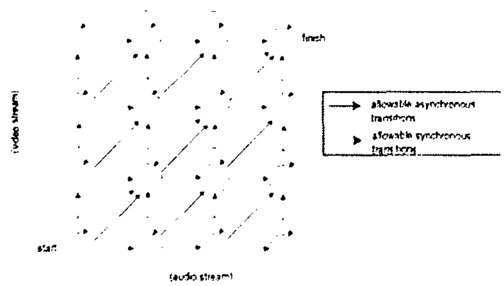
Fusion Techniques



Tsuhan Chen

Middle Integration (MI)

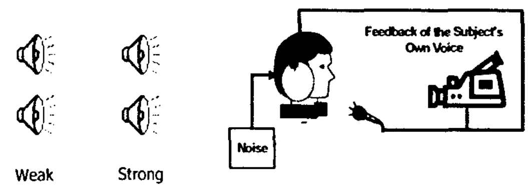
- Multistream HMM



Tsuhan Chen

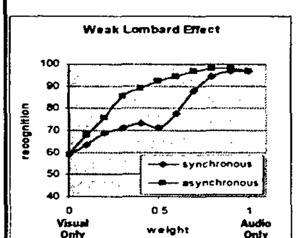
Lombard Effect

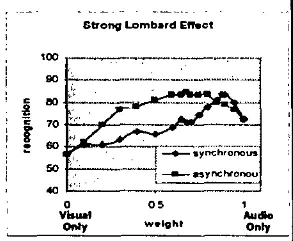
- Feedback
  - Voice changes with background noise
  - Lip movement changes too
- Data set



Tsuhan Chen

Result

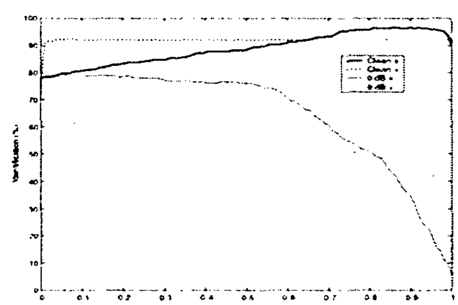




Tsuhan Chen

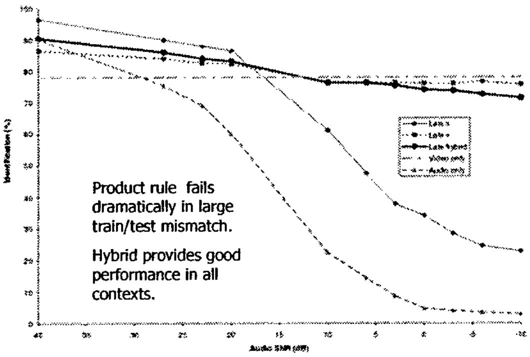
Product Rule vs. Sum Rule

(For Speaker Identification)



Tsuhan Chen

Product Rule vs. Sum Rule



Tsuhan Chen

Quick Recap

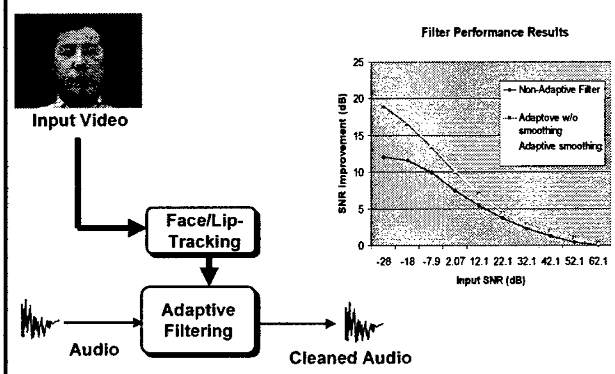
- Asynchronous MI (LI) has more freedom than synchronous MI (EI) → Better performance
- Product rule is better in Bayesian sense, but sum rule is more robust to mismatch
- Robustness to weighting
- Need to be careful about Lombard Effect
- Key to multimodal fusion
  - To model dependency between audio and visual signals
  - To dampen independent audio and visual noises

Tsuhan Chen

Beyond Multimodal ASR...

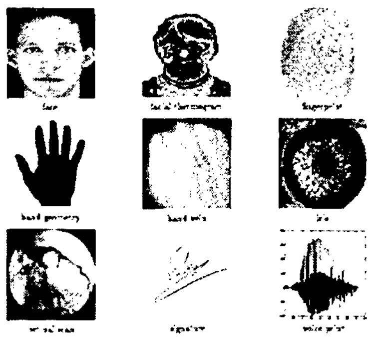
Tsuhan Chen

Visual-Assisted Speech Enhancement



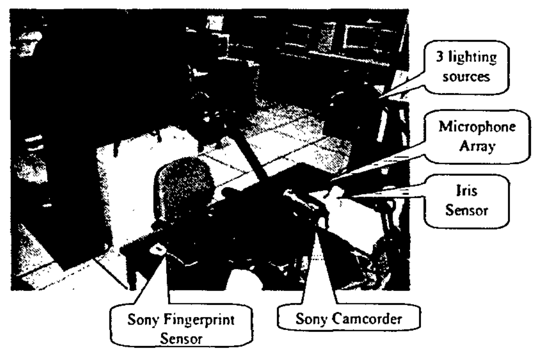
Tsuhan Chen

Multimodal Biometrics



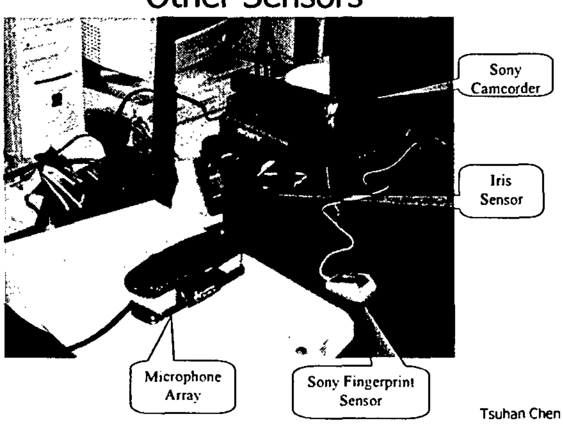
Tsuhan Chen

Data Collection



Tsuhan Chen

Other Sensors



Tsuhan Chen

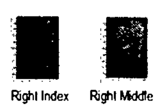
CMU Multimodal Biometrics Database

Face:

- 30 subjects with 300 images each
- Image size: 720*480
- Different lighting conditions, with/without glasses and ambient lighting

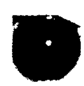
Fingerprint:

- Image size: 192*128
- 50 images each finger



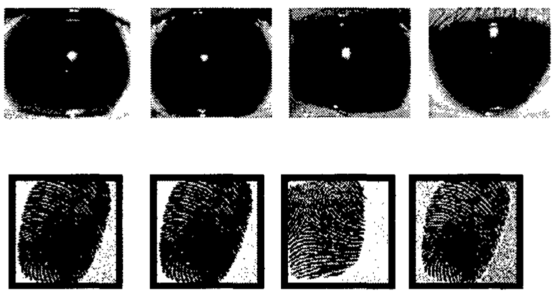
Iris

- Iris size: about 400*400
- 10 images each eye



Tsuhan Chen

Fingerprint and Iris Images



Tsuhan Chen

Multimodal User Interfaces

[CMU-GM Lab]

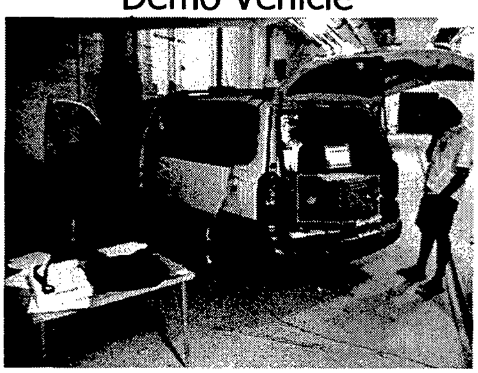
Face/Eye/Hand Tracking:
. Driver-Vehicle Interfaces
. Cognitive Overflow Study

Interview Video

Airbag Deployment Control
Mirror/wheel/panel/seat adjustment

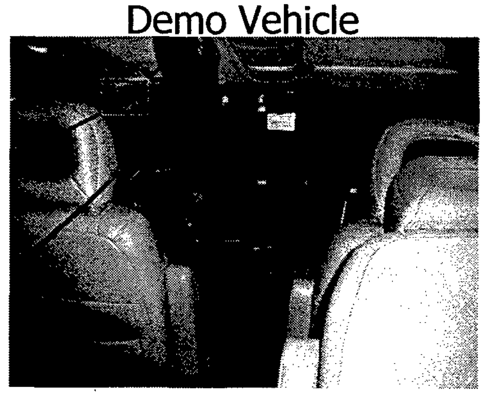
Driver ID and Encryption:
Security, Safety, User Preference

Demo Vehicle



Tsuhan Chen

Demo Vehicle



Tsuhan Chen

FaceCam/GestureCam

"Visual is not noise-free"

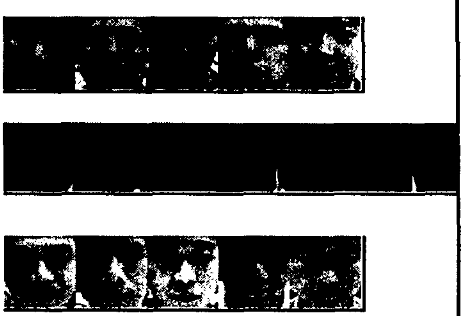
Tsuhan Chen

Challenges...

Pose/Registration

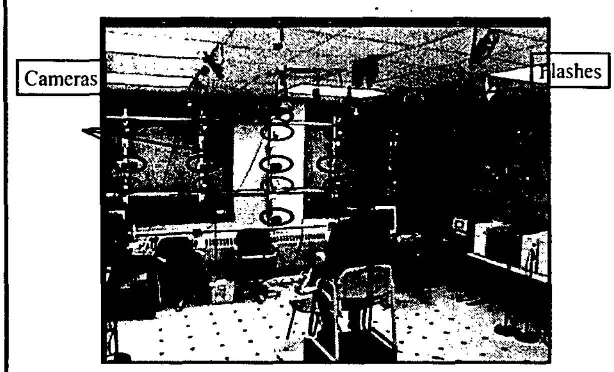
Illumination

Expression



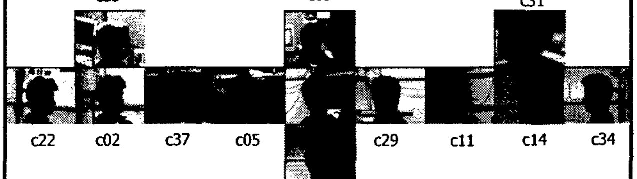
Tsuhan Chen

CMU PIE Database



Tsuhan Chen

Pose Variation



Tsuhan Chen

Illumination Variation

- 22 illumination conditions with background light

- 21 illumination conditions without background light

Tsuhan Chen

Conclusions

- Database is essential
  - Need to consider Lombard Effect
- Fusion is important
  - We can learn from acoustic ASR; we can perhaps lead ASR
- Confidence estimation is important
- Visual channel is not noise-free

Tsuhan Chen

Related Forums

- IEEE Multimedia Signal Processing (MMSP) Technical Committee, 1996~
- *Proceedings of IEEE*, Special Issue on MMSP, 1998
- IEEE MMSP Workshops
  - Princeton 1997, Los Angeles 1998, Copenhagen 1999, Cannes 2001, St. Thomas 2002
- IEEE International Conf. on Multimedia and Expo. (ICME)
  - New York 2000, Tokyo 2001, Lausanne 2002, Baltimore 2003
- *IEEE Transactions on Multimedia*, March 1999~
  - Special issues: networked multimedia 2001, multimedia database 2002, multimodal interface 2003

Tsuhan Chen

Advanced Multimedia Processing Lab

Please visit us at:

http://amp.ece.cmu.edu

Or, please email me at
tsuhan@cmu.edu

Tsuhan Chen